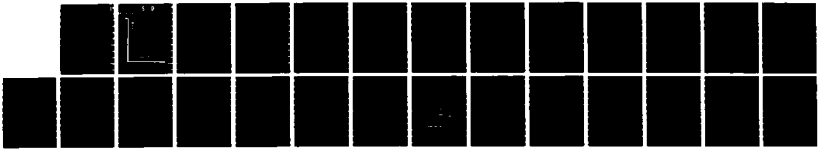
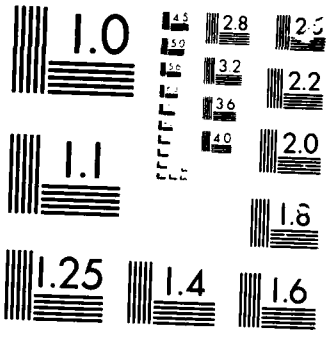
AD-A168 926 AIR FORCE OFFICER QUALIFYING TEST (AFOQT) RETESTING EFFECTS(U) AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX T O ARTH JUN 86 AFHRL-TP-86-8 1/1

UNCLASSIFIED F/G 5/9 NL





AFHRL-TP-86-8

AIR FORCE

HUMAN RESOURCES LABORATORY



# AIR FORCE OFFICER QUALIFYING TEST (AFOQT) RETESTING EFFECTS

Thomas O. Arth, 1Lt, USAF

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

June 1986

Interim Paper for Period December 1983 - January 1985





AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Scientific Advisor
Manpower and Personnel Division

RONALD L. KERCHNER, Colonel, USAF
Chief, Manpower and Personnel Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b. RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution is unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>AFHRL-TP-86-8 | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a. NAME OF PERFORMING ORGANIZATION<br>Manpower and Personnel Division<br>6b. OFFICE SYMBOL (If applicable)<br>AFHRL/MOAO | 7a. NAME OF MONITORING ORGANIZATION |
| 6c. ADDRESS (City, State, and ZIP Code)<br>Air Force Human Resources Laboratory<br>Brooks Air Force Base, Texas 78235-5601 | 7b. ADDRESS (City, State, and ZIP Code) |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>Air Force Human Resources Laboratory<br>8b. OFFICE SYMBOL (If applicable)<br>HQ AFHRL | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER |
| 8c. ADDRESS (City, State, and ZIP Code)<br>Brooks Air Force Base, Texas 78235-5601 | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. 62703F<br>PROJECT NO. 7719 7719<br>TASK NO. 18 18<br>WORK UNIT ACCESSION NO. 19 47 |

11. TITLE (Include Security Classification)
Air Force Officer Qualifying Test (AFOQT) Retesting Effects

12. PERSONAL AUTHOR(S)
Arth, Thomas O.

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT (Year, Month, Day) | 15. PAGE COUNT |
|---|---|---|---|
| Interim | FROM Dec 83 TO Jan 85 | June 1986 | 26 |

16. SUPPLEMENTARY NOTATION
This work was accomplished under TS Study Numbers 8524, 8749, and 8750.

| 17. COSATI CODES | | |
|---|---|---|
| FIELD | GROUP | SUB-GROUP |
| 05 | 09 | |
| 05 | 08 | |

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Air Force Officer Qualifying Test (AFOQT)
classification
selection tests
selection and classification
test-retest

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Retesting on the AFOQT is permitted within 6 months only with a waiver. This effort was conducted to determine the effects of retesting over various time intervals and to compare retesters with non-retesters. According to the results of t-tests and regression analyses, those who retest at less than 6 months benefit most from retesting. Also, retesters are a highly self-selected group. Further research is indicated in which subjects would be randomly assigned to retake the AFOQT over various time intervals. Implications for AFOQT retesting policy are discussed.

| | |
|---|---|
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS | 21 ABSTRACT SECURITY CLASSIFICATION<br>Unclassified |
| 22a NAME OF RESPONSIBLE NDIVIDUAL<br>Nancy A. Perrigo, Chief, STINFO Office | 22b TELEPHONE (Include Area Code)<br>(512) 536-3877<br>22c OFFICE SYMBOL<br>AFHRL/TSR |

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

AFHRL Technical Paper 86-8 June 1986

# AIR FORCE OFFICER QUALIFYING TEST (AFOQT) RETESTING EFFECTS

Thomas O. Arth

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

Douglas K. Cowan
Chief, Officer Selection and Classification Function

Submitted for publication by

Dr. Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

This publication is primarily a working paper. It is published solely to document work performed.

## SUMMARY

The Air Force Officer Qualifying Test (AFOQT) is a paper-and-pencil aptitude battery. Test results are used to make selection decisions based on Verbal (V) and Quantitative (Q) composite scores and classification decisions based on Pilot (P) and Navigator-Technical (N-T) composite scores. Retests are not permitted until after 6 months, unless the applicant can show the first testing did not reflect his/her true ability. A relatively large number of waivers of the 6-month requirement are granted. This study addressed the benefits of retesting by comparing retesters with non-retesters and by determining the effects of retaking the AFOQT over various time intervals.

Subjects were applicants for officer training who tested on Form O of the AFOQT between October 1981 and December 1983. This included 2,246 retesters and 42,776 non-retesters. The retesters were divided into four groups who were retested (a) in less than 6 months, (b) from 6 to 11 months, (c) from 12 to 17 months, and (d) after 18 months. T-test results indicated that retesters' initial scores were significantly lower than those of non-retesters and that they differed significantly among groups defined on the basis of time interval between retest. Regression analyses were performed to determine whether the four retest groups showed differing score gains. Retest scores were higher than initial test scores for all groups on all composites. The groups differed in amount of gain in P and N-T but not on the V and Q composites. The less-than-6-months group showed the largest gain, followed by the 6-to-11-months group. The 12-to-17-months group showed the least gain.

It was concluded that candidates who obtain a waiver benefit most by retesting, especially those applying for pilot and navigator training. Whether these findings stem from the candidates' having a valid reason for a waiver or from learning effects is not clear. Further research is needed to clarify time and composite effects associated with AFOQT retesting. However, practice effects would be minimized by allowing retests only after 12 months.

## PREFACE

This work was completed under Task 771918, Selection and Classification Technologies, which is part of a larger effort in Force Acquisition and Distribution. It was subsumed under work unit number 77191847, Development and Validation of Civilian and Non-rated Officer Selection Methodologies. This work unit was established in response to Air Force Regulation (AFR) 35-8, Air Force Military Personnel Testing System.

I want to express my thanks to the personnel of the Technical Services Division, especially Jim Brazel, Jim Friemann, and Cal Fresne, without whose help this paper would not have been possible.

TABLE OF CONTENTS



LIST OF FIGURES



LIST OF TABLES

# AIR FORCE OFFICER QUALIFYING TEST (AFOQT)
# RETESTING EFFECTS

## I. INTRODUCTION

The purpose of this research and development (R&D) effort was to investigate the effects of retesting on the Air Force Officer Qualifying Test (AFOQT). The AFOQT is a paper-and-pencil aptitude test battery used to make selection and classification decisions for Air Force officers. It was of special interest to determine the effects of retaking the AFOQT in less than 6 months. Air Force Regulation (AFR) 35-8, Air Force Military Personnel Testing System, dated March 1978, stated that an individual may not retest on the AFOQT in less than 4 months. However, in April 1983, this regulation was revised to increase the retesting restriction to 6 months. A retest is now permitted in less than 6 months if officially requested through the Major Command Test Control Officer (MAJCOM TCO) to the Air Force Military Personnel Center (AFMPC/MPCYPT) and approved. Approval of this waiver depends upon whether an individual can provide justification suggesting that the results of the first administration of the AFOQT did not reflect his/her true abilities. Examples of a valid reason for a waiver include illness and a recent death in the family. Because there was some ambiguity regarding the optimal time interval between the initial test and the retest, this investigation was undertaken to determine the effects of retesting over various time intervals.

The consistency of aptitude retest scores depends on (a) the extent to which aptitude changes and (b) test reliability. In theory, retesting on an aptitude test should result in no changes in scores if there were no changes in the underlying aptitude and the test was perfectly reliable. However, it is naive to assume that environmental influences do not affect individuals between tests to cause changes in aptitude. Humphreys (1978) even suggested that neither aptitude nor achievement should be used as labels on any test because of a disparity between theory and practice.

In fact, there is evidence that suggests scores on aptitude tests such as the AFOQT will improve over time due to both internal and external factors. A study was conducted by Christal (1984) in which he administered the AFOQT to members of an Air Force Reserve Officer Training Corps detachment every other year for 4 years. It was shown that the greatest gains were in the spatial subtests. These increases were followed by those in the numeric subtests while the verbal subtests showed the least amount of gain.

Other studies dealing with the effects of retesting on aptitude tests have compared the scores of retesters with those who do not retest. Givner, Klintberg, and Hynes (1980) examined the effects of retesting on the Medical College Admission Test by comparing retesters with non-retesters. They found that while there was some improvement in scores for retesters, their initial and retest scores were significantly less than those of examinees who did not retest. Similar findings were reported by Alderman (1981) for the Scholastic Aptitude Test (SAT).

Another problem of measuring change in aptitude tests is the reliability of the testing instrument. According to Cronbach and Furby (1970), any change in test performance, as measured by subtracting a pretest score from a posttest score, could lead to fallacious conclusions. The reason is that this change is systematically related to error of measurement. That is, individuals who score low initially would tend to score higher on any subsequent test whereas high scorers would tend to score lower. This tendency is called regression toward the mean. However, problems in interpreting measures of change may be avoided by taking into account standard error of measurement. The phenomenon of regression toward the mean and how to deal with it is discussed more fully by Cohen and Cohen (1975).

This investigation took two approaches to the problem of retesting. First, the retesters were divided into samples that had 6-month increments between tests. In this manner, maturation and/or learning effects could be studied along with the effects of waiving the 6-month restriction prescribed in AFR 35-8. Secondly, retesters' scores were compared with scores of those who did not retest. This was done to determine whether the average scores of retesters differed from the average scores of those who did not retest. It is likely that individuals who retested represent a self-selected group and therefore can be expected to be different from non-retesters.

## II. METHOD

The subjects were examinees tested on AFOQT Form O between October 1981 and December 1983. The subjects included examinees who retested on the AFOQT during this time as well as those who tested only once. Further, only Officer Training School (OTS) candidates were included as subjects. About 9% of these subjects were females, and approximately 80% had at least a college education and were between the ages of 21 and 27. Since the purpose of this R&D was to analyze differences among individuals retested below 6 months, those who retested above the 6 month point, and non-retesters, the subjects were assigned to the following samples:

Sample $R_{1-5}$ - Individuals who retested less than 6 months after first test (N = 312).

Sample $R_{6-11}$ - Individuals who retested at least 6 months but less than 12 months after first test (N = 1,300).

Sample $R_{12-17}$ - Individuals who retested at least 12 months but less than 18 months after first test (N = 443).

Sample $R_{18-27}$ - Individuals who retested 18 months or more after first test (N = 191). None was retested more than 27 months after first test.

Sample $R_{6-27}$ - Individuals who retested 6 months or more after first test (N = 1934).

Sample NR - Individuals who did not retest (N = 42,776).

The variables used in the analysis were Short Battery scores on five composites derived from 16 subtests which make up the AFOQT. Short Battery scores are a subset of all the items in the AFOQT and were used for decision making prior to January 1984. The composites are Pilot, Navigator-Technical, Academic Aptitude, Verbal, and Quantitative. Two sets of scores were obtained on each subject except for the non-retesters. Composite scores were percentiles ranging from 1 through 99. For the purposes of this study, these variables were labeled as follows:

$P_1$ - Pilot composite score on first testing.
$P_2$ - Pilot composite score on second testing.
$N\text{-}T_1$ - Navigator-Technical composite score on first testing.
$N\text{-}T_2$ - Navigator-Technical composite score on second testing.
$AA_1$ - Academic Aptitude composite score on first testing.
$AA_2$ - Academic Aptitude composite score on second testing.
$V_1$ - Verbal composite score on first testing.
$V_2$ - Verbal composite score on second testing.
$Q_1$ - Quantitative composite score on first testing.
$Q_2$ - Quantitative composite score on second testing.

Statistical techniques used for the data analysis included independent and dependent t-tests for differences between means. T-tests for related means were computed to detect differences between the two administrations of the AFOQT. Independent t-tests were used to compare the means between samples. As 120 t-tests were computed, the accumulation of Type I error in performing multiple t-tests was a potential problem. This was avoided by adopting a stringent level of significance (.001).

Linear models analysis was also used to predict what scores the retesters would have received on the second administration if their initial scores were held constant between samples. Linear models analysis is a technique in which a full model is compared with a restricted model through the use of F-tests. If no significant differences are found between the full and the restricted models, the restricted model can predict the criterion as well as the full model and is therefore used. If significant differences are found, the full model must be used to predict the criterion. A complete explanation of this procedure may be found in Ward and Jennings (1973). A diagram showing the full and restricted models, as well as how the determination was made as to which models to use in this research, is shown in Appendix A.

## III. RESULTS

Table 1 and Figure B-1 show the mean AFOQT composite scores for each sample. Comparing sample $R_{1-5}$ and sample $R_{6-27}$, percentile scores were found to be higher when individuals retested 6 months or more after the first test, as opposed to less than 6 months after the first test. A second general trend is seen when one compares mean scores among retest samples that were broken into 6-month increments. Sample $R_{1-5}$ generally had the lowest mean scores of all samples on both administrations. Means for sample $R_{6-11}$ were higher than those of sample $R_{1-5}$, but sample $R_{12-17}$ means were generally lower than those of sample $R_{6-11}$. The highest mean scores were found in sample NR while sample $R_{18-27}$ had the largest means among the retesters.

Table 1. Mean AFOQT Composite Scores

| | Samples | | | | | |
|---|---|---|---|---|---|---|
| | (N = 42,776) | (N = 312) | (N = 1,300) | (N = 443) | (N = 191) | (N = 1,934) |
| Composites | NR | $R_{1-5}$ | $R_{6-11}$ | $R_{12-17}$ | $R_{18-27}$ | $R_{6-27}$ |
| $P_1$ | 46.97 | 30.04 | 33.72 | 33.54 | 41.64 | 34.46 |
| $P_2$ | | 44.56 | 46.70 | 43.63 | 51.96 | 46.52 |
| $N\text{-}T_1$ | 46.59 | 27.66 | 31.28 | 30.32 | 39.74 | 31.90 |
| $N\text{-}T_2$ | | 40.83 | 43.28 | 39.76 | 49.19 | 43.06 |
| $AA_1$ | 47.67 | 25.28 | 28.08 | 28.14 | 38.41 | 29.11 |
| $AA_2$ | | 35.98 | 38.87 | 37.09 | 47.66 | 39.33 |
| $V_1$ | 52.47 | 33.21 | 35.34 | 36.26 | 45.00 | 36.51 |
| $V_2$ | | 43.06 | 44.47 | 44.30 | 52.79 | 45.25 |
| $Q_1$ | 44.49 | 24.26 | 27.23 | 26.25 | 36.30 | 27.90 |
| $Q_2$ | | 34.06 | 37.31 | 34.10 | 44.59 | 37.29 |

Independent t-tests were computed among the means of all samples shown in Table 1. The results, reported as level of significance obtained, are presented in Table 2. As shown in the first five rows of the table, mean scores for non-retesters were significantly higher than those for retesters in all samples on both the initial and second administration of the AFOQT. Two exceptions were noted in the second administration of the Pilot composite and in sample $R_{18-27}$. Only in sample $R_{12-17}$ was $P_2$ significantly lower than the non-retesters' Pilot score. The other exception was that the mean scores of sample $R_{18-27}$ on all composites of the second administration did not differ from those obtained by sample NR.

Table 2. Level of Significance Between Samples' Mean AFOQT Scores

| Sample Comparisons | AFOQT Composites | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $P_1$ | $P_2$ | $N\text{-}T_1$ | $N\text{-}T_2$ | $AA_1$ | $AA_2$ | $V_1$ | $V_2$ | $Q_1$ | $Q_2$ |
| NR vs. $R_{1-5}$ | .001 | .088 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 |
| NR vs. $R_{6-11}$ | .001 | .704 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 |
| NR vs. $R_{12-17}$ | .001 | .008 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 |
| NR vs. $R_{18-27}$ | .003 | .012 | .001 | .185 | .001 | .996 | .001 | .872 | .001 | .959 |
| NR vs. $R_{6-27}$ | .001 | .449 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 |
| $R_{1-5}$ vs. $R_{6-11}$ | .006 | .175 | .007 | .119 | .024 | .050 | .165 | .394 | .024 | .031 |
| $R_{1-5}$ vs. $R_{12-17}$ | .029 | .627 | .093 | .569 | .053 | .526 | .090 | .523 | .196 | .982 |
| $R_{1-5}$ vs. $R_{18-27}$ | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 |
| $R_{1-5}$ vs. $R_{6-27}$ | .001 | .208 | .002 | .150 | .002 | .022 | .029 | .173 | .005 | .029 |
| $R_{6-11}$ vs. $R_{12-17}$ | .881 | .028 | .420 | .011 | .955 | .169 | .492 | .905 | .395 | .015 |
| $R_{6-11}$ vs. $R_{18-27}$ | .001 | .008 | .001 | .003 | .001 | .001 | .001 | .001 | .001 | .001 |
| $R_{12-17}$ vs. $R_{18-27}$ | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 | .001 |

Next, comparisons focused on examinees who retested in less than 6 months (sample $R_{1-5}$). As shown in the middle part of Table 2, two trends emerged from comparisons among sample $R_{1-5}$ and the other three retest samples. Generally, scores for sample $R_{1-5}$ were lower on the initial administration than those of other retesters. However, on the second administration, mean scores for examinees who retested in less than 6 months were significantly lower than those of only examinees who retested in greater than 18 months.

The final set of comparisons reported in Table 2 were made among the three retest samples to whom the AFOQT was readministered at the 6-month point or later. Only the comparison between samples $R_{6-11}$ and $R_{12-17}$ showed no significant differences in the initial administration of the AFOQT on all five composites. However, both samples scored significantly lower on most AFOQT composites on the initial and second administration than did sample $R_{18-27}$.

Table 3 shows the mean increases in composite scores for all samples. In every case, retest means were considerably higher than original test means. The largest increases generally occurred in sample $R_{1-5}$. Across all samples, the mean increase was the greatest for the Pilot and Navigator-Technical composites. The test-retest correlations for all retesters on each composite were as follows: Pilot = .812, Navigator-Technical = .852, Academic Aptitude = .853, Verbal = .880, and Quantitative = .775. These reliabilities were lower than expected, most likely due to the restricted variability in the sample.

Table 3. Mean Percentile Increase of AFOQT Scores Between First and Second Administrations

| Sample | N | Pilot | Nav-Tech | Academic Apt | Verbal | Quantitative |
|---|---|---|---|---|---|---|
| $R_{1-5}$ | 312 | 14.52 | 13.17 | 10.71 | 9.85 | 9.80 |
| $R_{6-11}$ | 1300 | 12.98 | 12.00 | 10.79 | 9.13 | 10.09 |
| $R_{12-17}$ | 443 | 10.09 | 9.44 | 8.95 | 8.03 | 7.85 |
| $R_{18-27}$ | 191 | 10.32 | 9.45 | 9.25 | 7.79 | 8.29 |
| $R_{6-27}$ | 1934 | 12.05 | 11.16 | 10.22 | 8.74 | 9.40 |

Note. Some of these figures vary slightly from Table 1 due to rounding, but all increases were significant at the .001 level.

Since the majority of samples showed significant differences among their initial scores, it became necessary to compute regression analyses to determine what their second scores would have been if their first scores had been the same. By doing this, the results could be analyzed as though each sample were drawn from the same population.

These analyses revealed curvilinear relationships on all five composites between the initial score and the second predicted score (see Appendix B). Group effects were found in the Pilot, Navigator-Technical, and Academic Aptitude composites but not in the Verbal and Quantitative composites. An interaction effect was apparent only with Academic Aptitude. The regression models found to be significant for the five composites were as follows: Model 3 for Pilot and Navigator-Technical, Model 1 for Academic Aptitude, and Model 2 for Verbal and Quantitative (see Appendix A).

The increase of scores at the 25th, 50th, and 75th percentiles, as predicted from the regression analysis, is shown in Table 4. It should be noted that a noticeable difference in predicted increases exists for those samples who retested in less than 12 months versus those who retested 12 months or more after the first administration. The largest predicted increases of the Pilot, Navigator-Technical, and Academic Aptitude composites occurred for samples $R_{1-5}$ and $R_{6-11}$. Another finding was that the largest predicted increases were observed below the 50th percentile across all composites. This was entirely expected.

Table 4. Predicted Percentile Score Increases on AFOQT Composites by Retest Interval

| AFOQT Percentile | Retester Samples | | | |
|---|---|---|---|---|
| | $R_{1-5}$ | $R_{6-11}$ | $R_{12-17}$ | $R_{18-27}$ |
| | | Pilot | | |
| 25 | 16.5 | 14.9 | 12.2 | 13.2 |
| 50 | 15.9 | 14.3 | 11.6 | 12.7 |
| 75 | 8.5 | 7.0 | 4.3 | 5.3 |
| | | Navigator - Technical | | |
| 25 | 15.2 | 13.9 | 11.5 | 12.0 |
| 50 | 15.4 | 14.1 | 11.7 | 12.1 |
| 75 | 8.4 | 7.1 | 4.7 | 5.2 |
| | | Academic Aptitude | | |
| 25 | 10.9 | 12.5 | 10.4 | 10.4 |
| 50 | 9.6 | 13.0 | 10.0 | 10.9 |
| 75 | 7.9 | 6.6 | 4.6 | 7.0 |
| | | Verbal | | |
| 25 | 10.6 | 10.6 | 10.6 | 10.6 |
| 50 | 10.3 | 10.3 | 10.3 | 10.3 |
| 75 | 5.6 | 5.6 | 5.6 | 5.6 |
| | | Quantitative | | |
| 25 | 11.1 | 11.1 | 11.1 | 11.1 |
| 50 | 8.1 | 8.1 | 8.1 | 8.1 |
| 75 | .8 | .3 | .3 | .3 |

## IV. DISCUSSION

All retesting produced significant increases in subjects' AFOQT scores. It then became interesting to speculate on what influences caused the increase in scores. Possible causes include regression to the mean, maturation, learning (i.e., practice effects and coaching), or different motivations for retesting. Each of these will be discussed in turn.

Regression to the mean cannot explain the magnitude of score increase observed in the retesters' scores. The AFOQT is a highly reliable test instrument with reliabilities ranging from .689 to .922 across the 16 subtests. Although the standard error of measurement (5.92) explains most of the change in the smallest mean percentile score increase (7.79), there are still other factors which may account for the differences.

It is highly unlikely that maturation could have caused the score increases in all samples. If that were the case, the score increases would have been greater as the time interval between tests increased. Additionally, it is doubtful that maturation could have occurred in less than 6 months. Maturation may have been a factor with the score increases in sample $R_{18-27}$, however. This was the only sample whose second AFOQT scores equaled the non-retesters' scores.

When examining the effects of learning, two areas need to be considered. One is the practice effect of having recently taken the test, and the other is the effect of possible coaching between tests. In a study by Johnson, Flinn, and Tyer (1979), it was shown that spatial skills significantly improve with practice. The data in the present study showed the greatest gains in the Pilot and Navigator-Technical composites, which are largely composed of spatial tests. Furthermore, this effect was most pronounced with those subjects who retested in less than 12 months. Verbal and Quantitative scores were not as susceptible to change. A less likely cause of the increase in scores would be coaching. DerSimonian and Laird (1983) reported small but positive effects of coaching on SAT scores. That is, they changed true scores by teaching subject matter, not testwiseness tricks. Since the subjects in this study were as likely to be coached before the initial administration of the AFOQT as between administrations, this probably was not the cause of the increase.

Motivation for retaking the AFOQT may have caused the increase in scores. Because the retesters had relatively low scores, their motivation surely was to increase their scores. The motivation to retake the AFOQT in sample $R_{18-27}$ may have differed from the other retester samples. According to AFR 35-8, individuals are required to retest if their scores are more than 2 years old and they are applying for commissioning or flying training. Therefore, sample $R_{18-27}$ score increases may have been caused either because of their differing motivation or because they matured between tests.

The regression analysis accomplished to equate the initial scores on all samples supported the contention that learning did occur in the Pilot and Navigator-Technical composites. Some slight maturation effects were also shown in these composites in that sample $R_{18-27}$ posted higher retest scores than did sample $R_{12-17}$. No group differences were found in the Verbal and Quantitative composites, which leads to the conclusion that these are the most stable of the aptitude indicators.

When the results of the regression analysis were compared with the obtained mean increase of scores in Table 3, two observations were made. First, the predicted increases corresponded with the obtained increases in the Pilot and Navigator-Technical composites. That is, the largest increases were found in sample $R_{1-5}$, followed by samples $R_{6-11}$, $R_{18-27}$, and $R_{12-17}$. Furthermore, there was a noticeable difference in score increases in these composites between those who retested above and below 12 months. Second, the predicted increase in scores was a

function of not only time between test administrations but also initial score level. That is, greater score increases may be expected at the 25th percentile than the 75th percentile. This was particularly relevant because the mean initial scores for all retesters across all of the composites were below the 50th percentile.

When mean scores of sample NR were compared with mean scores of all of the retester samples, the results showed that sample NR generally scored significantly higher than the retesters on both administrations. This indicates that personnel who decided to retest probably did so to increase their scores. Despite the increase of their scores, they would continue to be discernible from non-retesters, who posted higher scores.

The only exceptions to this finding would be in explaining the data from sample $R_{18-27}$ and the second Pilot administration. Sample $R_{18-27}$ may have had a different motive in retaking the AFOQT in that they probably did so to keep their scores current. $P_2$ scores were not different from sample NR's Pilot scores, which seems to indicate that some learning occurred in the Pilot composite subtests.

## V. CONCLUSIONS

Two questions were addressed in this study. One concerned the effects of retesting over various time intervals. The other was whether retesters were similar to non-retesters. The following findings were obtained.

First, regardless of the time interval between administrations of the AFOQT, increases occurred. In the case of waiving the 6-month retesting restriction, whether the increase was due to learning or being valid cases for the waiver is debatable. The regression analysis showed that the highest gains were found for samples $R_{1-5}$ and $R_{6-11}$. Moderate increases were also found for sample $R_{18-27}$, which indicated that some slight maturation effects possibly occurred. However, if a goal of retesting is to minimize the effects of practice, then the minimum time to allow retesting is 12 months after the first administration. This is especially critical for those individuals applying for pilot or navigator training.

Also, the linear models analysis revealed that the amount of gain depends not only on time between tests but also on initial score. Since most retesters score low initially, large improvements in AFOQT scores may be expected. However, only in marginal cases would retaking the AFOQT substantially improve an individual's chances of being selected into OTS given the competitive nature of today's recruiting environment.

Finally, individuals who retested were a highly self-selected group. Although retesters' scores were improved by taking the AFOQT again, they still did not equal non-retesters' scores, and retesters were discernible from non-retesters. Therefore, even though large increases in scores may be expected with retesting, those increases would not be sufficient in most cases to change an applicant's chances of being selected into OTS. However, given the large increases in Pilot scores, especially of those who retested in less than 12 months, pilot classification decisions may be altered by retesting.

Future research is indicated from these results. In a follow-up study, subjects should be randomly assigned to four groups after initial administration of the AFOQT. One group would retest shortly after the first test. The other groups would retest 6, 12, and 18 months later. This paradigm would control different motivation factors for retesting (i.e., illness on the first test, keeping test scores current) while measuring learning and maturation effects.

## REFERENCES


Alderman, D. (1981, Win). Student self-selection and test repetition. Educational and Psychological Measurement, 41, 1073-1081.

Christal, R. (1984). Texas A&M retest study. Unpublished manuscript. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L., & Furby, L. (1970). How we should measure "change" - or should we? Psychological Bulletin, 74, 68-80.

DerSimonian, R., & Laird, N. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. Harvard Educational Review, 53(1), 1-15.

Givner, N., Klintberg, I., & Hynes, K. (1980). Effect of retaking the Medical College Admission Test on applicants' scores. Psychological Reports, 47, 411-415.

Humphreys, L. (1978). Relevance of genotype and its environmental counterpart to the theory, interpretation, and nomenclature of ability measures. Intelligence, 2, 181-193.

Johnson, S., Flinn, J., & Tyer, Z. (1979). Effect of practice and training in spatial skills on embedded figures scores of males and females. Perceptual and Motor Skills, 48, 975-984.

Ward, J., & Jennings, E. (1973). Introduction to linear models. Englewood Cliffs, NJ: Prentice-Hall.

## APPENDIX A: SPECIFICATIONS FOR MULTIPLE LINEAR REGRESSION ANALYSIS

Table A-1. Model Specifications

| Model | Component Predictors |
|---|---|
| 1 | $Y' = U + G_1 + G_2 + G_3 + G_4 + Apt + Apt^2 + G_1Apt + G_1Apt^2 + G_2Apt + G_2Apt^2 + G_3Apt + G_3Apt^2 + G_4Apt + G_4Apt^2$ |
| 2 | $Y' = U + Apt + Apt^2$ |
| 3 | $Y' = U + G_1 + G_2 + G_3 + G_4 + Apt + Apt^2$ |
| 4 | $Y' = U + G_1 + G_2 + G_3 + G_4 + G_1Apt + G_2Apt + G_3Apt + G_4Apt$ |
| 5 | $Y' = U + G_1 + G_2 + G_3 + G_4 + Apt$ |
| 6 | $Y' = U + Apt$ |
| 7 | $Y' = U + G_1 + G_2 + G_3 + G_4$ |

Note. These seven models were run for each of the five composites.

$Y'$ - predicted second score

U - unit vector

$G_1$ - membership in Sample $R_{1-5}$ coded 1 if a member; Ø otherwise

$G_2$ - membership in Sample $R_{6-11}$ coded 1 if a member; Ø otherwise

$G_3$ - membership in Sample $R_{12-17}$ coded 1 if a member; Ø otherwise

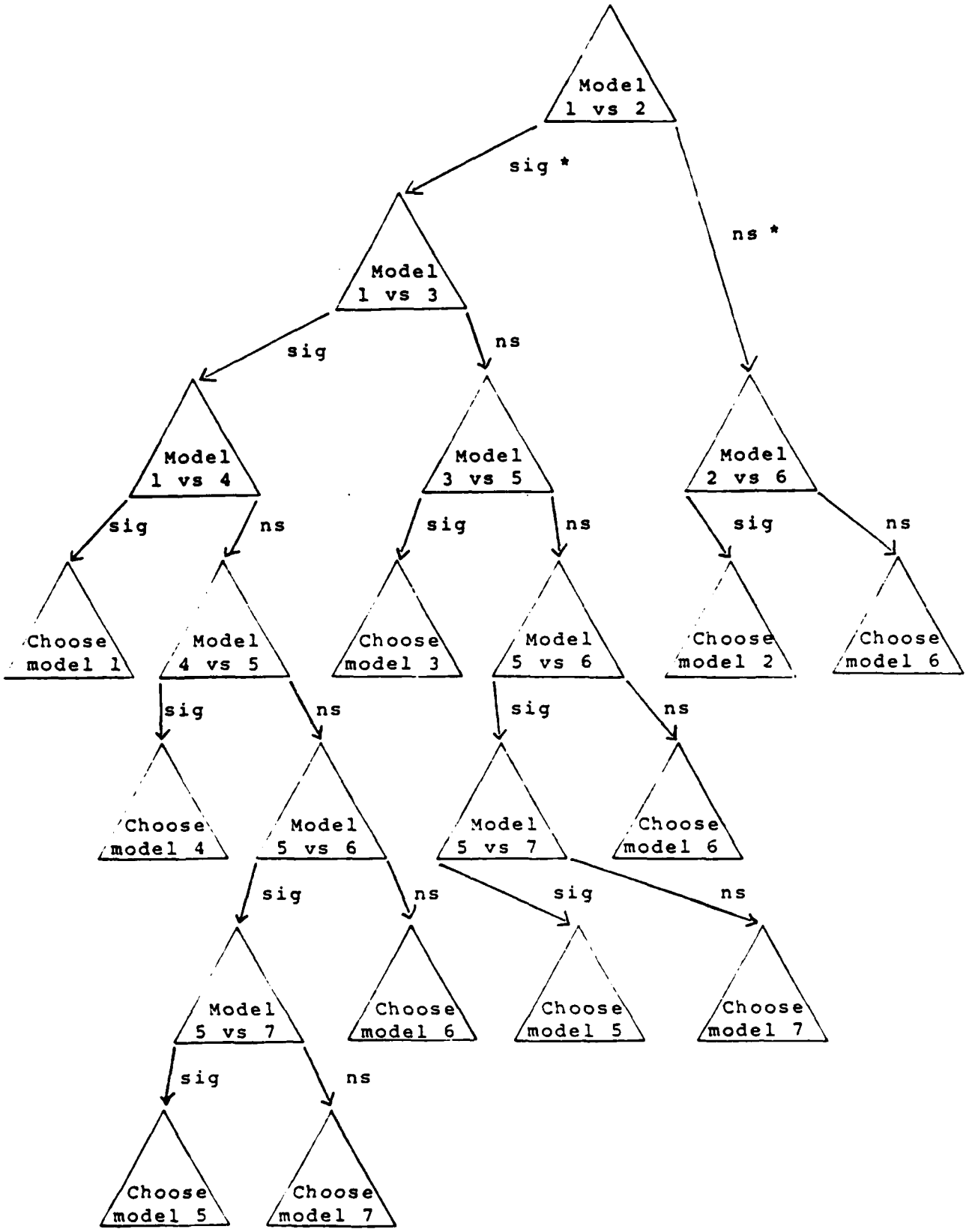
$G_4$ - membership in Sample $R_{18-27}$ coded 1 if a member; Ø otherwise

Apt - one of the five composite scores

$Apt^2$ - composite score squared

$G_{1-4}Apt$ - interaction term

$G_{1-4}Apt^2$ - squared interaction term




* sig = significant; ns = not significant.

Figure A-1. Sequential F-test Comparisons.

APPENDIX B: CURVILINEAR RELATIONSHIPS ON FIVE COMPOSITES
INITIAL VERSUS RETEST SCORE COMPARISONS

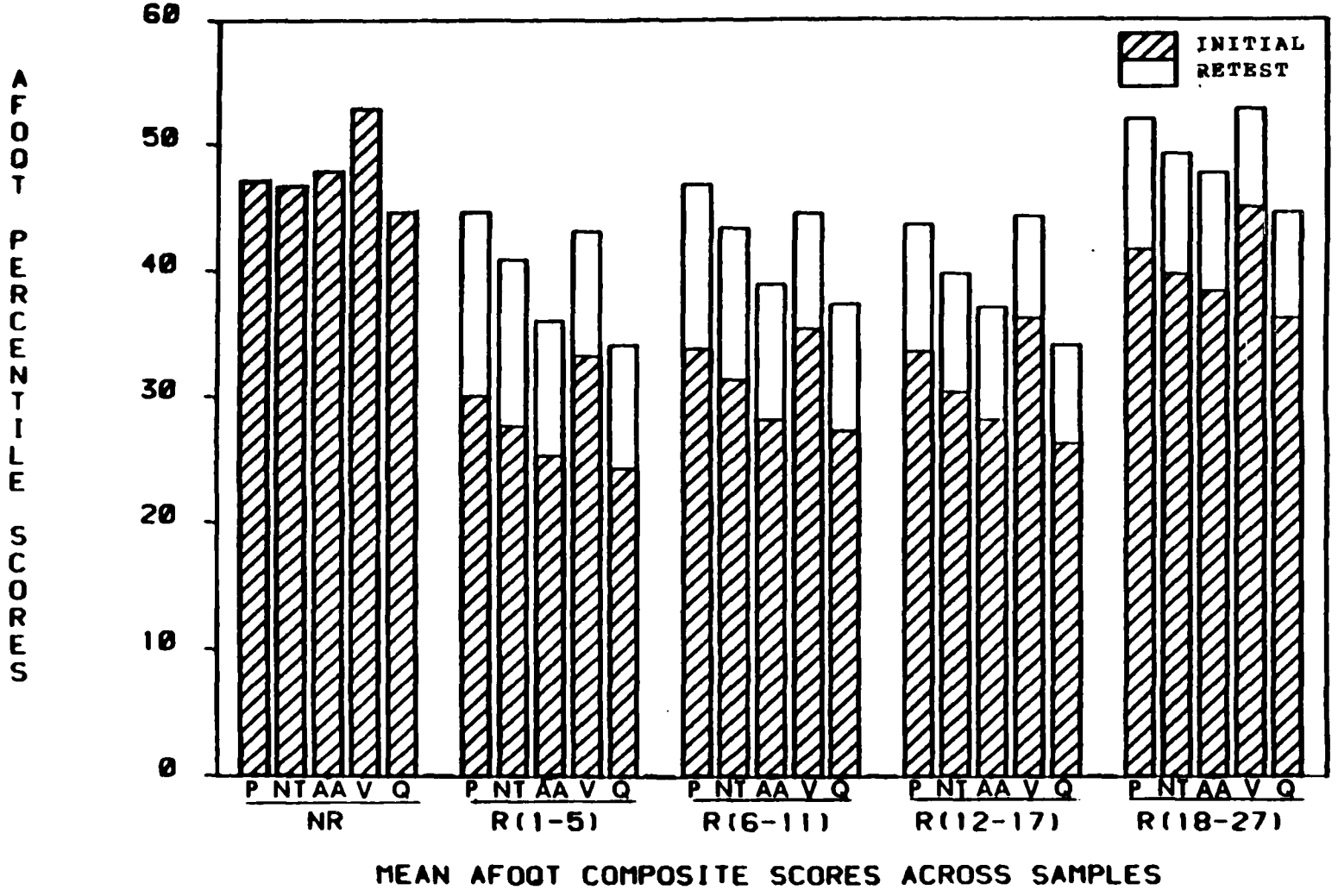



Figure B-1. Initial versus Retest Score Comparisons.

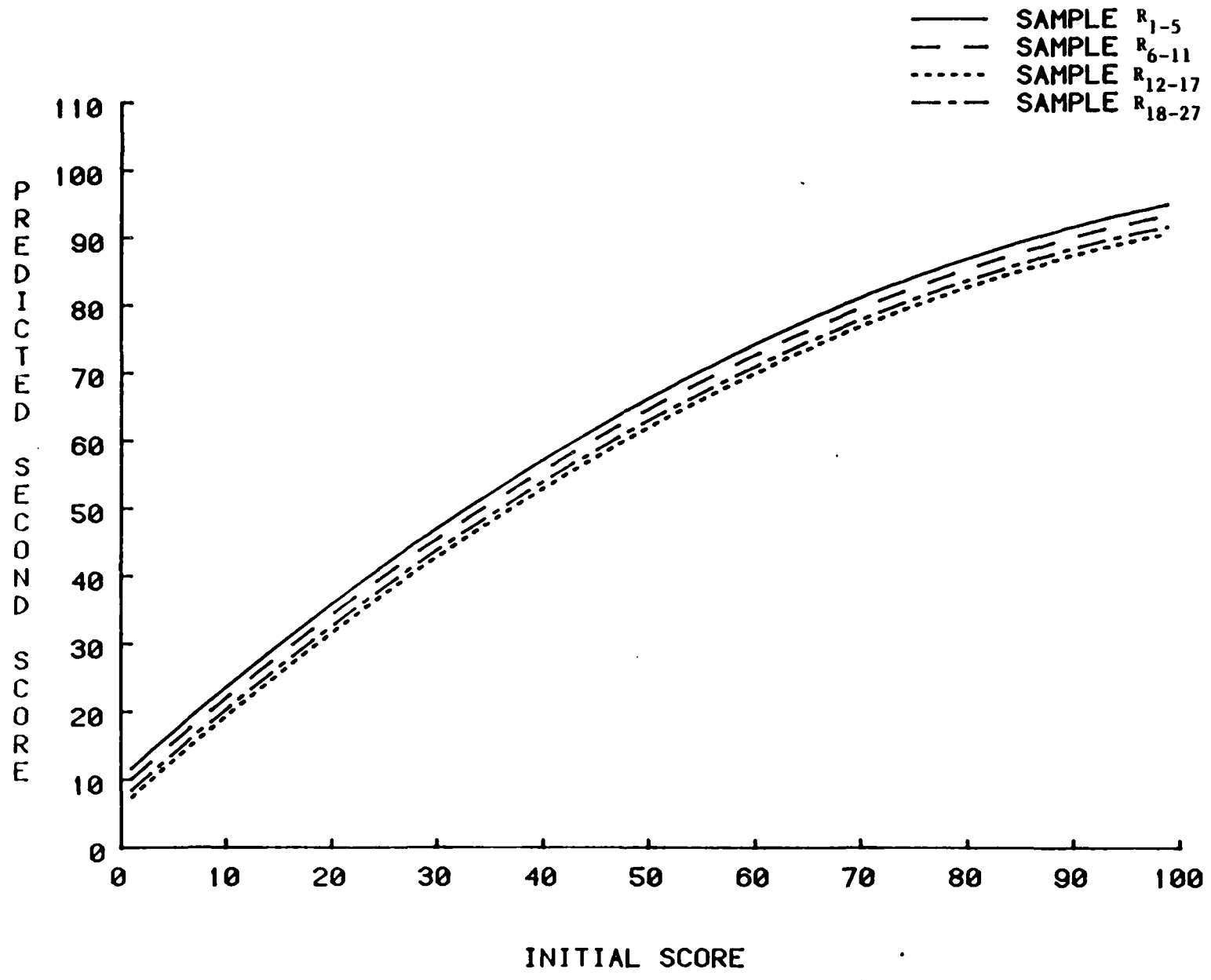



Figure B-2. Pilot Composite Regression Lines.

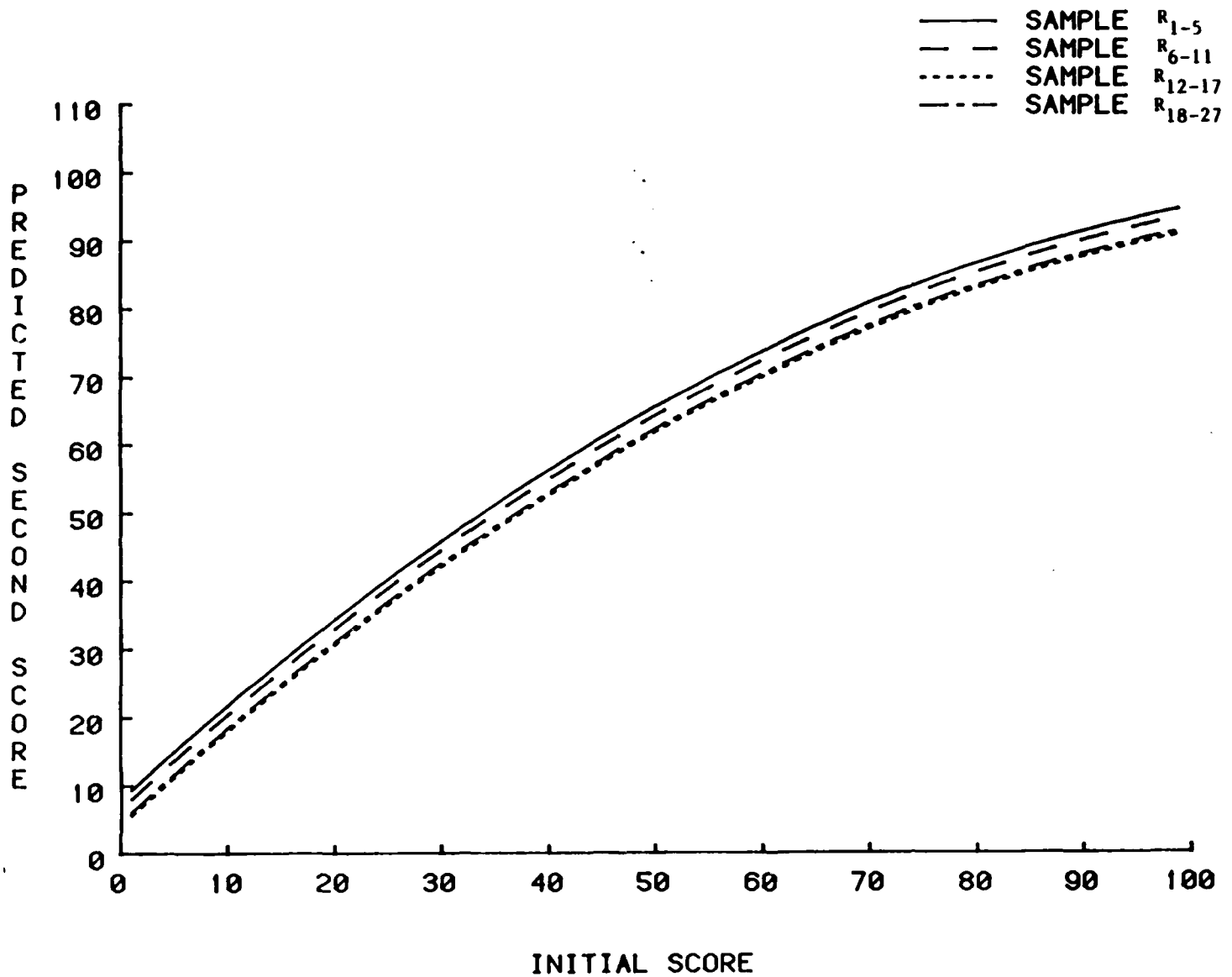



Figure B-3. Navigator-Technical Composite Regression Lines.

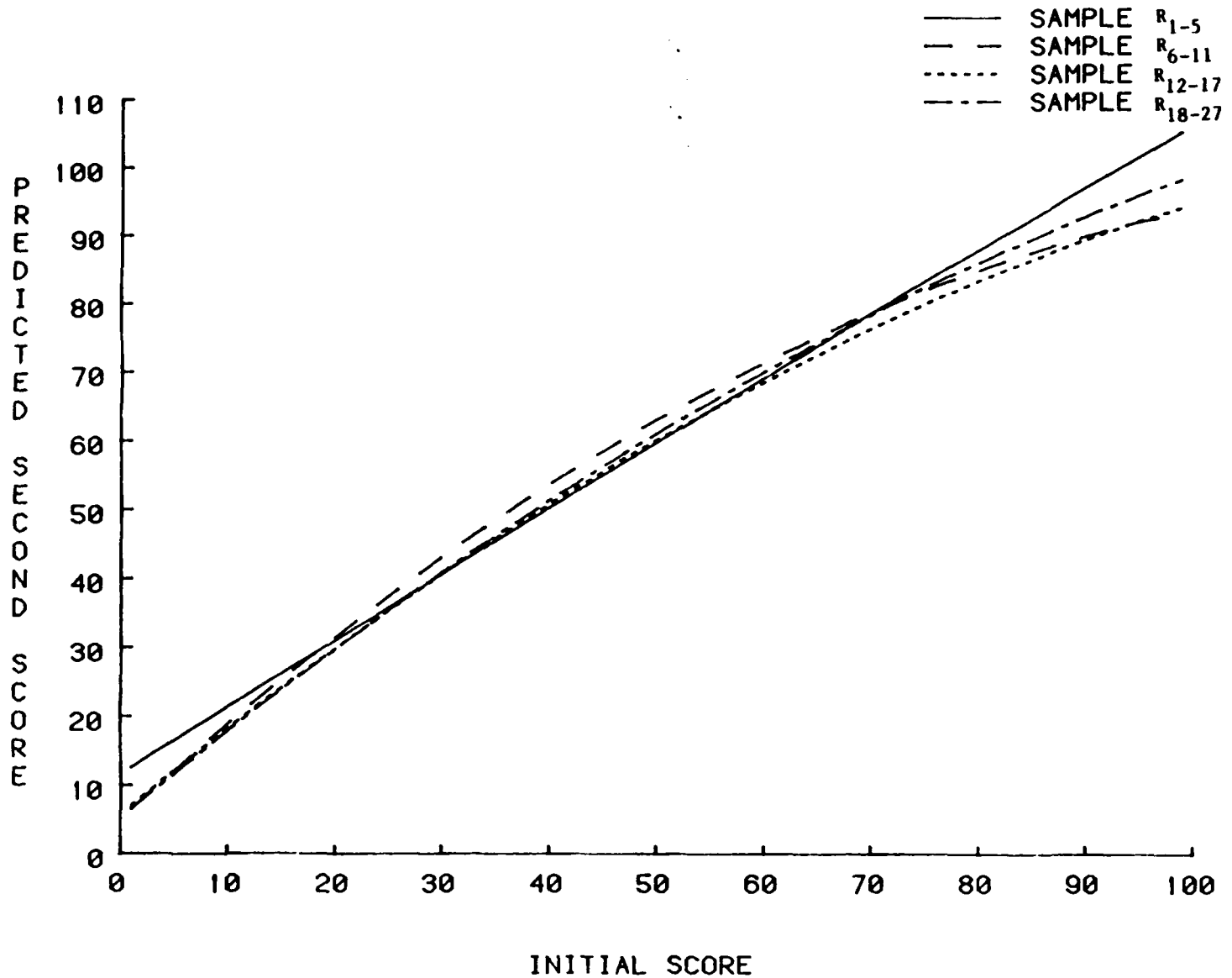



Figure B-4. Academic Aptitude Composite Regression Lines.

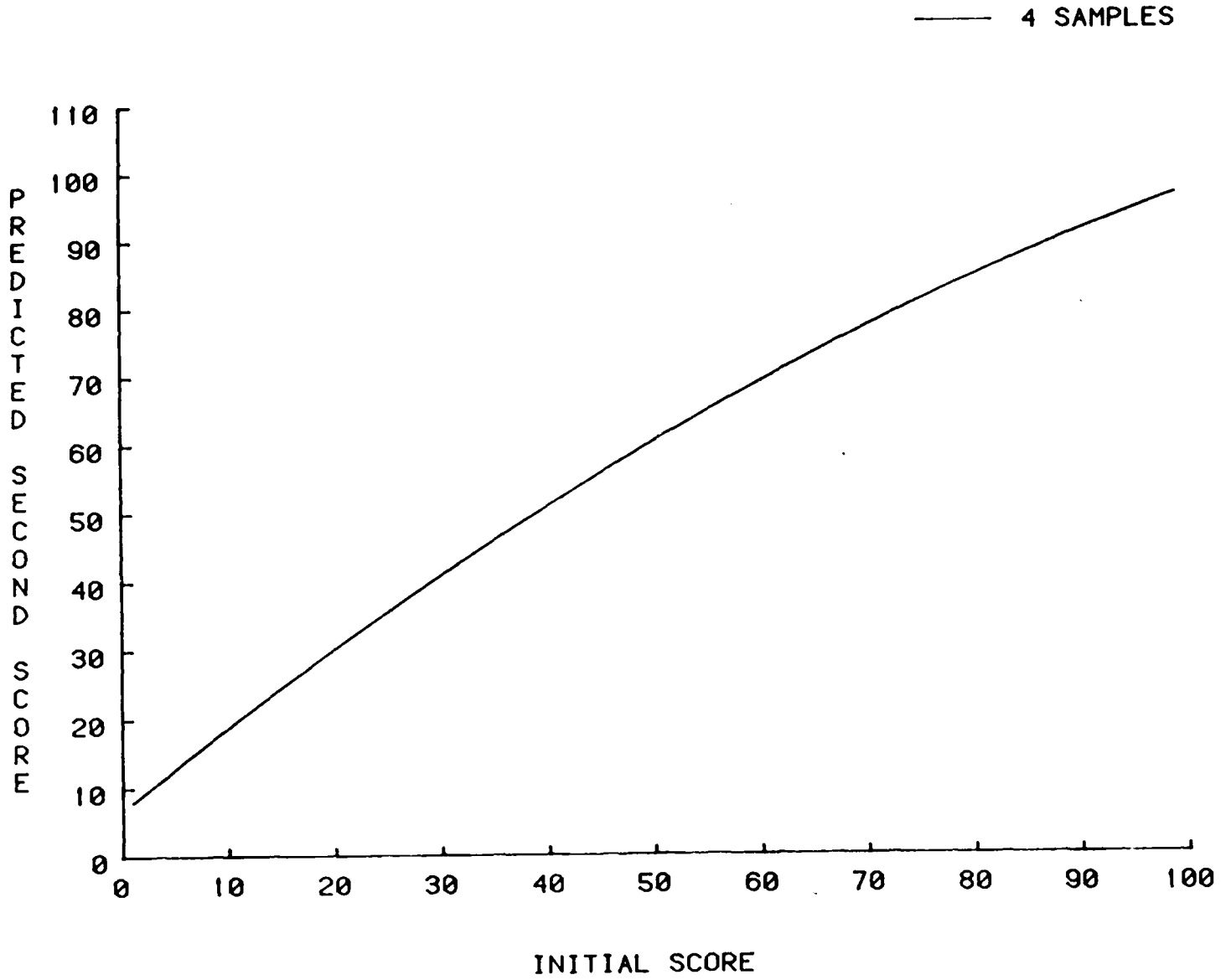



Figure 8-5. Verbal Composite Regression Line.

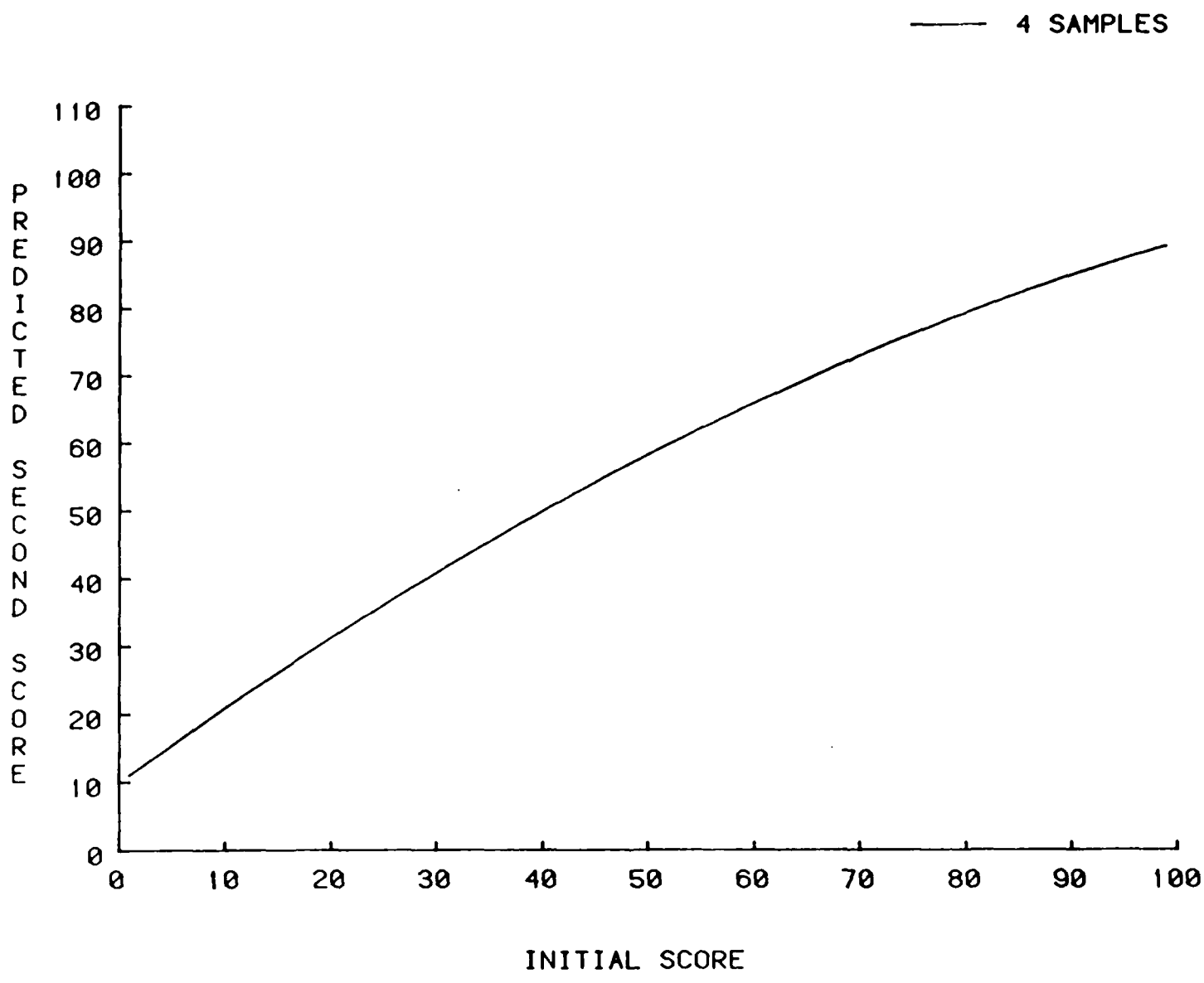



Figure B-6. Quantitative Composite Regression Line.

END

DTIC

7-86